Automatic generation and evaluation of sparse protein signatures for families of protein structural domains

MATTHEW J. BLADES,[1] JON C. ISON,[2] RANJEEVA RANASINGHE,[2] AND JOHN B.C. FINDLAY[3]

[1]AstraZeneca R&D Charnwood, Loughborough, Leicestershire LE11 5RH, England
[2]MRC Rosalind Franklin Centre for Genomics Research (formerly the MRC UK HGMP Resource Centre), Hinxton, Cambridgeshire CB10 1SB, United Kingdom
[3]School of Biochemistry and Molecular Biology, University of Leeds, Leeds, LS2 9JT, United Kingdom

(RECEIVED July 6, 2004; FINAL REVISION September 7, 2004; ACCEPTED September 7, 2004)

Abstract

We identified key residues from the structural alignment of families of protein domains from SCOP which we represented in the form of sparse protein signatures. A signature-generating algorithm (SigGen) was developed and used to automatically identify key residues based on several structural and sequence-based criteria. The capacity of the signatures to detect related sequences from the SWISSPROT database was assessed by receiver operator characteristic (ROC) analysis and jack-knife testing. Test signatures for families from each of the main SCOP classes are described in relation to the quality of the structural alignments, the SigGen parameters used, and their diagnostic performance. We show that automatically generated signatures are potently diagnostic for their family (ROC50 scores typically >0.8), consistently outperform random signatures, and can identify sequence relationships in the "twilight zone" of protein sequence similarity (<40%). Signatures based on 15%–30% of alignment positions occurred most frequently among the best-performing signatures. When alignment quality is poor, sparser signatures perform better, whereas signatures generated from higher-quality alignments of fewer structures require more positions to be diagnostic. Our validation of signatures from the Globin family shows that when sequences from the structural alignment are removed and new signatures generated, the omitted sequences are still detected. The positions highlighted by the signature often correspond (alignment specificity >0.7) to the key positions in the original (non-jack-knifed) alignment. We discuss potential applications of sparse signatures in sequence annotation and homology modeling.

Keywords: sparse protein signature; SCOP; domain; protein family; ROC analysis; EMBOSS

The limitations of using purely sequence-based methods to identify protein evolutionary relationships are well documented (Brenner et al. 1998; Rost 1999; Spang and Vingron 2001). Although developments in this field have shown significant improvement (Altschul et al. 1997; Karplus et al. 1998; Park et al. 1998), it is well known that the use of structural information can probe more distant relationships than sequence alone (Hargbo and Elofsson 1999; Ison et al. 2000; Blake and Cohen 2001; Jennings et al. 2001).

The relationship between the sequence of a protein and its three-dimensional structure is not strictly defined. This is particularly noticeable where highly dissimilar sequences adopt similar structures. This has led researchers to investigate, both experimentally and computationally, the idea that structural determinants are restricted to a limited number of "key residues positions" in the sequence (Friedberg and Margalit 2002). Mutation experiments have shown that many proteins retain their activity (Markiewicz et al. 1994; Suckow et al. 1996) and stability (Milla et al. 1994) despite the introduction of mutations at many positions in the se-

Reprint requests to: Matthew J. Blades, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, England; e-mail: matthew.blades@astrazeneca.com; fax: +44 (0) 1509-645557.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.04929005.

quence. Bowie et al. (1990) showed that the mutations that did have an effect on the structure or activity were located in functional sites or the hydrophobic core. The identity of these core residues did not appear to be crucial, and interchange between similar hydrophobic residues did not perturb the structure. The computational studies of Shakhnovich et al. (1996), Mirny et al. (1998), and Mirny and Shakhnovich (2001) reached similar conclusions.

Recent improvements in structure alignment and comparison methods have focused attention more toward the study of the "key" structural residues and the residue clusters in which they are involved (Artymiuk et al. 1994; Dosztanyi et al. 1997; Kannan and Vishveshwara 1999; Kleywegt 1999; Orengo 1999; Kannan et al. 2001; Reddy et al. 2001; Li et al. 2002). Collectively these studies show that the folding of very different sequences into similar three-dimensional structures is mediated by the interactions of a very small number of key residues at specific positions in the sequence. It is in this area of research that we developed our SIGNATURE approach (Daniel et al. 1999; Ison et al. 2000).

Our approach identifies structurally important residues in a protein family and incorporates them in a descriptor known as a sparse protein signature (Daniel et al. 1999). A signature is a sparse representation of a protein family (according to the SCOP hierarchy), consisting of residue identities within key residue positions (signature positions) and flexible gaps that represent the sequence positions between successive key residue positions (Fig. 1). A signature is suitable for scanning against a sequence database, from which homologous sequences can be identified, using the SIGSCAN program (which forms part of the EMBOSS suite) (http://www.hgmp.mrc.ac.uk/EMBOSS).

SIGSCAN allows flexibility in the alignment of a signature to a sequence. For example, when matching signature positions to residues any gap sizes are permitted, but gap penalties are incurred when gaps are used that do not appear in the signature. SIGSCAN scores residues using a residue substitution matrix (see Table 5). Each signature-sequence match is given a score of the best alignment of the signature positions to the protein sequence. The highest-scoring matches in a search of a sequence database will, ideally, belong to the family from which the signature was derived.

The SIGNATURE approach rests on two premises:

1. Although the majority of residues in proteins are subject to substitution, tight constraints are imposed on the evolution of a few key residues responsible for determining and maintaining the fold.

2. The identification of these key residues can yield signatures which when scanned against a sequence database can identify protein homologous relationships.

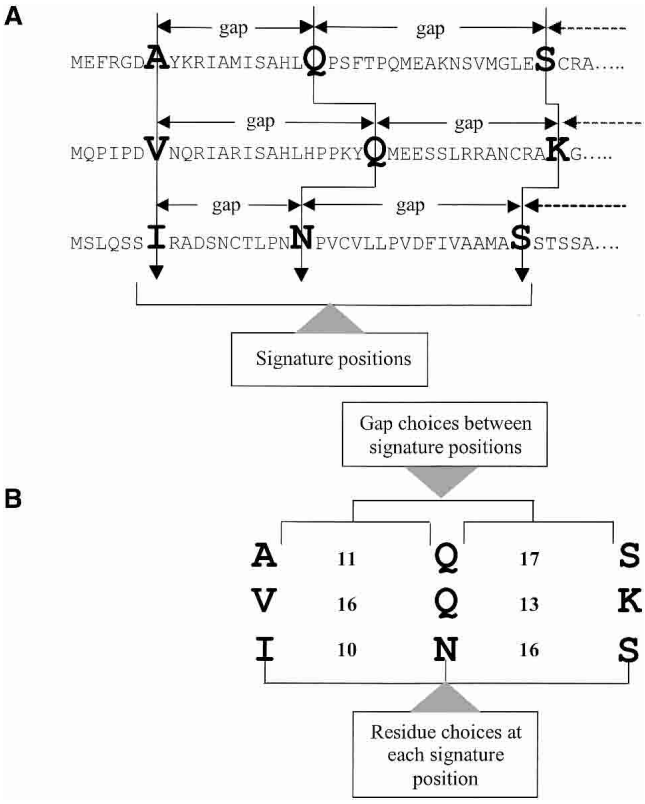

Figure 1. The construction of a sparse protein signature is illustrated. *A* shows three sequence segments; the residues in large text indicate the key residues selected from that region. *B* illustrates the data contained in a signature, i.e., the residue choices at each position and the gap choices between successive signature positions.

We have shown that very sparse signatures (i.e., those with a few positions only) are in fact diagnostic of protein families (Ison et al. 2000). This early approach was an essentially manual method; however, its automation should make it more widely useful and permit more rigorous validation. Here we describe the automation of the approach and the new SigGen program, which generates signatures automatically by applying sequence- and structure-based scoring schemes to "seed alignments" of protein families taken from the SCOP database (Lo Conte et al. 2000).

Test signatures for families from each of the main SCOP classes are discussed in relation to the quality of the seed alignments, the SigGen scoring scheme used, their diagnostic performance, and the quality of the resulting signature-sequence alignments. We were interested to know whether sparse signatures could incorporate sufficient information to identify more distantly related sequences than those found in a single SCOP family. To test this, signatures were generated for the "Globin-like" superfamily from SCOP. Such superfamilies contain domains whose sequence identities may be low, but whose structural and often functional features suggest a common evolutionary origin.

Preparation of seed alignments

Table 1 shows the protein families considered and their position in the SCOP hierarchy. Representatives were selected from each of the major SCOP classes. In order to assess the dependence (if any) of signature performance on the relatedness of the proteins in the seed alignment, two "seed sets" of structures were created for each test family such that within each set no pair of structures has greater than a threshold sequence similarity. Thresholds of 90% and 50% were used to create the "DATA90" and "DATA50" seed sets, respectively. The structure alignment program STAMP (Russell and Barton 1992) was used to generate structure-based sequence alignments from these sets (see Materials and Methods).

Generating signatures for the Globin-like superfamily required a single set of superfamily members. The superfamily contains three families: globin, phycocyanin, and truncated hemoglobin. The DATA50 seed sets for each family were combined and a seed alignment generated. STAMP, however, rejected the truncated hemoglobin family because the structures for this family did not achieve an alignment score above the required threshold (according to the STAMP scoring scheme) when aligned to the rest of the set. Thus the superfamily seed alignment contained structures from the phycocyanin and globin families only.

Signature generation

SigGen uses two sequence-based and two structure-based scoring schemes with the option to combine more than one scheme in the residue selection process. An alignment position is only scored (and therefore a candidate for inclusion in the signature) if the residues from the different domains are "structurally equivalent": This is defined by the STAMP algorithm (see Materials and Methods). Thus a signature position corresponds to a set of structurally equivalent residues occupying the same (high-scoring) position in the STAMP alignment.

The first sequence-based scheme (ResId) uses a residue substitution matrix (BLOSUM62). The score for an alignment position is the average residue substitution score for all possible pairs of substitutions at each structurally equivalent position. The second scheme (ResVar) applies the residue variability function of Mirny and Shakhnovich (2001) to the residues at each alignment position. The structure-based schemes do not consider residue identity at all. The first of these (N-Con) determines the total number of contacts for every residue. The score for an alignment position in this case is the average number of contacts made by each residue at that position. The score calculated in the C-Con scheme is the degree to which these contacts are conserved in the family. The scoring schemes are described in greater detail in the Materials and Methods.

The amount of information a signature contains depends on its sparsity. For example, a if a seed alignment contains eight domains with an average sequence length of 100 residues, then a signature of 10% sparsity will contain key residue and gap data from the 10 highest-scoring seed alignment positions. A signature of 20% sparsity will consider the top 20 positions, and so on. SigGen was used to generate signatures at sparsities of 2%, 5%, 10%, 15%, 20%, 25%, and 30% and with the four different scoring schemes to yield 28 signatures from the DATA90 data set and a further 28 signatures from the DATA50 data set for each of the test families. Twenty-eight signatures were also generated from the Globin-like superfamily seed alignment.

Evaluation of signature performance

Signatures were scanned against release 40 of the SWISSPROT database using the SIGSCAN program in EMBOSS. We evaluated the diagnostic performance of signatures using receiver operator characteristics (ROC) analysis. This has been used for many years in clinical studies to evaluate the usefulness of diagnostic tests, for example, serum cholesterol level as a diagnostic for heart disease (Gribskov and Robinson 1996). A ROC curve plots sensitivity ("rate of

Table 1. *SCOP families considered*

Class	Fold	Superfamily	Family	Gold standard
All α	DEATH domain	DEATH domain	DEATH domain	25
All α	Globin-like	Globin-like	Globin	820
All β	Lipocalin	Lipocalin	Fatty acid binding protein-like	89
All β	Lipocalin	Lipocalin	Retinol binding protein-like	116
All β	Viral coat and capsid proteins	Viral coat and capsid proteins	Plant virus proteins	21
α/β (mixed)	TIM β/α barrel	Triosephosphate isomerase (TIM)	Triosephosphate isomerase (TIM)	109
α/β (mixed)	TIM β/α barrel	(Trans) glycosidases	Type II chitinase	49
α + β (segregated)	Lysozyme-like	Lysozyme-like	C-type lysozyme	89

Protein families were taken from the SCOP database. The Class, Fold, Superfamily, and Family classification of the families are given. The number of sequences from SWISSPROT that were inferred to be related to each family is given under "Gold standard" (see Materials and Methods).

true positives") on the *Y*-axis against (1−specificity) ("rate of true negatives") on the *X*-axis, calculated for all rank positions in a list of hits that is rank-ordered on the basis of score (highest score first). The area under a ROC curve can be calculated and is a measure of the probability of correct classification. For example, in the case of signatures, an area of 0.88 indicates that a sequence belonging to the same family as that from which the signature was derived has a probability of 0.88 of scoring higher than a sequence known to be unrelated to the signature. In practice, ROC curves are usually truncated to the first 50 or 100 false hits and the area calculated to generate a ROC50 or ROC100 score. These scores are quicker and more convenient to calculate, can be expressed by fewer decimal places, and reflect the way in which the biologist, who is not prepared to search through large numbers of false hits, will use the method. We use the ROC50 score here. We also quote signature sensitivity, which is defined as the proportion of the total number of known true hits that are detected before the 50th false hit. In calculating the ROC50 score and sensitivity, we use a "gold standard" of known protein family members as described in Materials and Methods.

Validation of signatures

A good test of the diagnostic performance of signatures and also the signature concept is to generate random signatures from the seed alignments and compare their performance to signatures generated by using a scoring scheme. SigGen provides an option to randomly select residues from anywhere in the alignment; this option was used here. As a further validation, jack-knife ("miss one out") testing was performed for the Globin family by removing one structure at a time from the seed set, rebuilding the alignment, and generating a new jack-knifed signature. Each domain in the Globin family DATA50 seed set was omitted in turn to generate a set of jack-knifed signatures at each of the seven sparsities used previously (2%, 5%, 10%, 15%, 20%, 25%, and 30%). These were scanned against the SWISSPROT database, and the list of hits returned by the search was checked for the presence of the omitted seed domain. A seed domain was defined as "detected" if it scores higher than the 50th false hit. The diagnostic performance of the jack-knifed signatures was assessed by ROC analysis as before. We also assessed the impact of jack-knifing on the detection of distantly related family members, with a view to understanding whether domains in the seed set exert equal influence on the signature. The quality of the signature-sequence alignments that result from searching a signature against SWISSPROT were investigated by calculating alignment specificity scores (see Materials and Methods) for the jack-knifed and original (all structures) signatures. Our aim was to determine whether the residues highlighted in the alignment of a signature to a SWISSPROT sequence were indeed structurally equivalent to those in the original seed alignment from which the signature was generated.

Results

Signature performance

Table 2 shows the sparsity and ROC50 score of the best-performing signature derived from the DATA90 seed set, for each SigGen scoring scheme and test family. Here we describe each test family in turn and then the Globin family in more detail. The Chitinase and lysozyme family signatures from each scoring scheme all achieved perfect or near-perfect ROC50 scores (1.0). This is not particularly surprising, because the sequences in the seed alignments are closely related. The DEATH domain family signature produced consistently poor ROC50 scores, owing to the poor quality of the seed alignment. In fact, the most recent release of the SCOP database now subdivides the original DEATH domain family into three separate families, Caspase Recruitment Domain (CARD), DEATH Domain (DD), and DEATH Effector Domain (DED). The FABP family has a good seed alignment, and the signatures produce consistently high ROC50 scores for each scoring scheme. The Globin, Plant Virus, and RBP families are unusual cases in that the ROC50 scores for the different schemes are identical. This is because the seed alignments contain a limited number of structurally equivalent residues. Therefore, less sparse signatures than those in Table 2 could not be generated, because there were insufficient structurally equivalent residues. Therefore signatures generated using the different scoring schemes will contain the same key residues and will achieve the same ROC50 scores. The Plant Virus family signatures performed poorly. These are Jelly-Roll proteins with some variations in structure between family members. Some proteins have an additional 1–2 β-strands, and this variation compromises the alignment

Table 2. *Best-performing signatures from each family*

Family	ResId BLOS 62	C-Con	N-Con	ResVar	Average ROC50 score
Chitinase	1.00 (20)	1.00 (20)	1.00 (20)	1.00 (15)	1.00
DEATH	0.50 (15)	0.37 (30)	0.45 (30)	0.56 (20)	0.47
Globins	0.84 (15)	0.84 (15)	0.84 (15)	0.84 (15)	0.84
Plant virus	0.45 (15)	0.45 (15)	0.45 (15)	0.45 (15)	0.45
FABP	0.90 (15)	0.89 (25)	0.89 (30)	0.90 (30)	0.90
Lysozyme	0.97 (5)	0.97 (15)	0.97 (15)	0.97 (5)	0.97
TIM	0.77 (15)	0.76 (15)	0.77 (30)	0.77 (20)	0.77
RBP	0.75 (20)	0.75 (20)	0.75 (20)	0.75 (20)	0.75

The best-performing signatures (i.e., those with highest ROC50 scores) derived from the DATA90 seed sets. The columns represent the four different scoring schemes, and the numbers in the columns are the ROC50 scores, with the signature sparsity in parentheses. The different scoring schemes are described in the Materials and Methods section.

and therefore the signature. The RBP and globin family seed alignments were also poor; however, this was not due to structural differences, but to the fact that the DATA90 seed set contained large numbers of structures. Nonetheless, both signatures were still able to achieve significant ROC50 scores (>0.75).

Figure 2 displays graphically the distribution of signature sparsity amongst the best signatures (Table 2). Signatures of sparsity of 15% and 20% occurred most frequently among the best-performing signatures; however, no single SigGen scoring scheme consistently outperforms the others. Signatures of sparsity of 40%, 50%, and 60% were also tested (data not shown) and found to be progressively less discriminating. These signatures include positions which are not structurally equivalent, or evolutionarily conserved and which are unlikely to be key to the family. Their inclusion therefore incorporates nonspecific information ("noise"), with a resulting decrease in diagnostic performance.

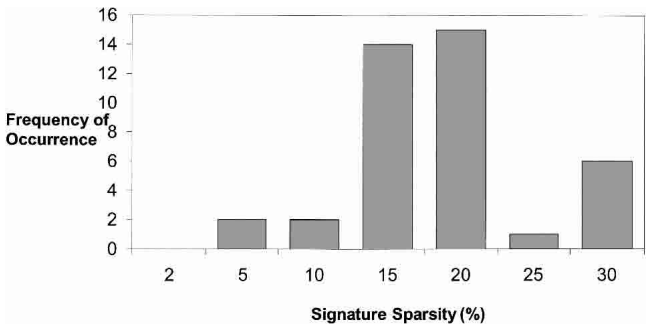

Figure 2. The graph displays the distribution of signature sparsity for the best-performing signatures derived from the DATA90 set for each of the SigGen scoring schemes and test families (see Table 3). A peak is visible at 15%–20% sparsity.

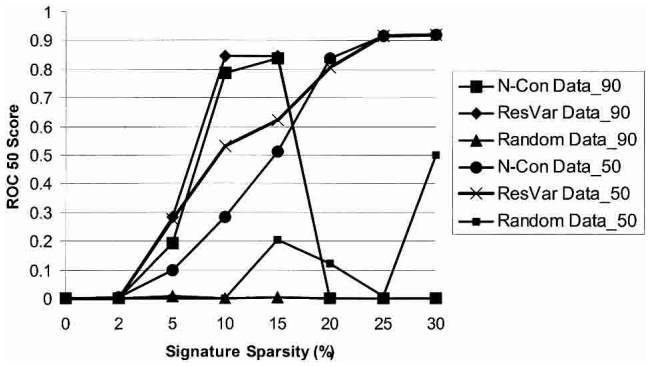

Figure 3. The graph shows the ROC50 scores plotted against signature sparsity for the N-Con, ResVar, and random signatures for the Globin family derived from the DATA90 and DATA50 seed sets. The difference in performance between random and nonrandomly generated signatures is clear. Signature sparsity refers to the number of signature positions contained in each signature and is a percentage of the average sequence length for all domains in the seed alignment.

Globin family signatures (DATA90 set)

Figure 3 shows ROC50 values plotted against signature sparsity for the N-Con, ResVar, and random globin signatures. The 15% N-Con signature is discriminating with a ROC50 score of 0.84 with 88% of the known true relatives detected before the 50th false hit. Initially almost all hits detected are TRUE, followed by a sudden increase in FALSE and UNKNOWN hits around hit number 700. Similar results were obtained for the C-Con and ResId globin signatures (data not shown). The randomly generated signatures perform extremely poorly; only two signatures (5% and 15%) detected any true hits at all, and less than 1% of the true hits were detected by both of these.

The globin family DATA90 seed set included 44 structures with an average pairwise sequence similarity of 55%. However, owing to the large number of structures (as was the case for the RBP family), the alignment contained few regions of structural equivalence, resulting in signatures with a maximum sparsity of 15%.

We generated a matrix of sequence similarity data from an all-versus-all comparison of the TRUE hits retrieved by the 15% N-Con signature to the seed sequences. The TRUE hits have a sequence similarity to the 44 structures in the alignment that ranges from 30% to 99% with an average pairwise similarity of 55%. Thus, from an alignment of many structures and with extensive regions of poor residue equivalence, a sparse signature can be generated automatically that provides good coverage of the globin family, including the more diverse relatives.

Globin family signatures (DATA50 set)

The globin DATA50 seed set contained four structures with an average pairwise sequence similarity of 42%. The alignment quality is dependent to an extent on the number of structures present, and in this case the fewer structures yielded a better alignment than the DATA90 set. All of the N-Con and ResVar globin signatures above 15% sparsity perform well, with the ROC50 score peaking at 0.91 and 0.92, respectively, for the 30% signatures, indicating excellent discrimination of the globin family. Both the N-Con and ResVar 25% and 30% sparse signatures have a sensitivity of over 90%.

The random signature shows a peak ROC50 score of 0.5 (see Fig. 3) which, although higher than the random signatures for the DATA90 set, is still much worse than the signatures generated by using the scoring schemes. It appears that when alignment quality is high the choice of residues is still important, but is not so decisive as when alignment quality is poor (DATA90 set), where the informed (nonrandom) selection of key residues is essential to

producing a discriminating signature. We conclude therefore that the quantity of information contained in the DATA90 sets was unnecessary. Not only were superior alignments obtained with the DATA50 seed sets, but in the case of the globin family, the performance of the signatures increases. It is noticeable however that the 10% and 15% N-Con DATA90 signatures were much more discriminating than the equivalent DATA50 signatures. More positions must be sampled from the alignment containing fewer domains to achieve a similar level of performance.

The sequence similarity of the TRUE globin hits to the seed proteins varies from 28% to 98%. Signatures encoding a subset of residues generated from a set of four structures (42% average pairwise similarity) contain sufficient information to detect almost 800 globin sequences (including many divergent examples) and produce high (>0.9) ROC50 scores. This supports the idea that a sparse signature of key residues for a protein family can be generated automatically.

Jack-knife testing for Globin family

All of the N-Con and ResVar globin signatures of 5% sparsity and above generated from the non-jack-knifed DATA50 seed alignment detected all four of the seed proteins from the SWISSPROT database before the 50th false hit. This is hardly surprising, but is significant considering that random signatures failed to detect the jack-knifed seeds in more than 50% of cases. The Globin DATA50 seed alignment contains four domains, producing four different jack-knifed seed alignments. Signatures of sparsity >10% generated from these using the four SigGen scoring schemes always identify the omitted seed protein before the 50th false hit, but some of the more sparse signatures (2%, 5%, 10%) failed to do so. None of the random jack-knifed signatures detected any omitted seed proteins at all.

Table 3 shows the ROC50 scores for the best performing signatures generated using the N-Con, ResVar, and Random scoring schemes from the four jack-knifed and one non-jack-knifed alignments; similar results were obtained for the ResId and C-Con scoring schemes (data not shown). The N-Con and ResVar SigGen scoring schemes appeared to perform similarly in terms of detecting jack-knifed seeds and consistently outperform the random signatures. It is clear that omission of the domains d1ash_, d1d8ua_, or d1hlb_ does not have a significantly detrimental effect on signature performance. Removal of domain d1dxtb_, however, does. The sensitivity dropped from 92% to 75%, and there was an increase in the number and rank of the false hits with the ROC50 dropping from ~0.9 to 0.5. This highlights how critical the information obtained from the d1dxtb_ domain is in producing a discriminating signature. This result is explained by considering the sequence similarity between the TRUE hits detected and the seed proteins.

Table 3. *Optimum ROC50 scores for jack-knifed signatures*

SigGen scoring scheme	SCOP domain identifier of jack-knifed seed protein				Non jack-knifed results
	d1ash__	d1d8ua_	d1dxtb_	d1hlb__	
N-Con	0.91	0.91	0.50	0.90	0.92
ResVar	0.90	0.88	0.52	0.86	0.92
Random	0.02	0.28	0.01	0.02	0.50

The best ROC50 scores for N-Con, ResVar and random signatures of various sparsities derived from structural alignments, each of which is missing a different structure. The structure that was omitted is given at the top of the column. The far right column shows the best ROC score for signatures derived from the complete alignment.

The average pairwise sequence similarity between the domains d1ash_, d1d8ua_ and d1hlb_, and the TRUE hits detected ranges between 40% and 45%, with individual pairwise values as low as 28%. The average pairwise sequence similarity between domain d1dxtb_ and the TRUE hits detected is 66.3% with the lowest being 37%. Thus d1dxtb_ is a better representative of the family as a whole than the other three structures and thus has a powerful effect on diagnostic performance. This also supports the second premise on which the approach is based, that is, that key residue positions can define residue and gap information that when incorporated into a signature are diagnostic for SCOP families. Clearly, the choice of domains used in the structural alignment can have a significant impact on the signatures performance, and the information from a single domain can be critical.

In all jack-knifed data sets (including the poor-performing d1dxtb_ set), there are examples of the detection of divergent relationships (i.e., TRUE hits with average pairwise sequence similarities to the seed proteins of <40%). For example in the data set with the domain d1hlb_ removed, the protein with primary accession number P02224 (a Globin precursor from *Chironomas thummi thummi*, the midge) is detected and has pairwise similarity values to the domains d1d8ua_, d1ash_, and d1dxtb_ of 39.10%, 38.40%, and 38.10%, respectively.

Evaluation of signature–sequence alignments

Table 4 shows the sparsity and alignment specificity scores for the jack-knifed N-Con and ResVar signatures. C-Con and ResId signatures produced similar results (data not shown). Alignment specificity scores represent the accuracy of the signature-to-sequence alignments (see Materials and Methods). Alignment specificity scores of at least 0.74 were obtained for signatures generated from each of the jack-knifed seed alignments using the N-Con and ResVar scoring schemes, with many examples obtaining perfect (1.0) or near-perfect scores. Some of the lower-sparsity signatures did not perform as well, though, with scores of <0.50. The

Table 4. *Alignment specificity scores for jack-knifed signatures*

Domain jack-knifed from alignment	SigGen scoring scheme used to generate signatures			
	N-Con		ResVar	
	Sparsity	Specificity	Sparsity	Specificity
d1hlb__	5	—	5	—
	10	0.81	10	—
	15	0.83	15	0.95
	20	0.87	20	0.97
	25	0.85	25	0.97
	30	1.0	30	0.93
d1dxtb_	5	—	5	—
	10	—	10	—
	15	—	15	—
	20	—	20	—
	25	—	25	—
	30	—	30	0.81
d1d8ua_	5	—	5	—
	10	—	10	—
	15	0.43	15	—
	20	0.74	20	0.71
	25	0.74	25	0.74
	30	0.47	30	0.57
d1dash__	5	—	5	—
	10	—	10	—
	15	0.87	15	—
	20	0.89	20	0.94
	25	0.91	25	0.96
	30	0.91	30	0.94

For each jack-knifed data set, the sparsity, scoring scheme, and alignment specificity score for the signatures that were able to detect the jack-knifed seed proteins are shown. Dashes indicate where signatures failed to detect the jack-knifed protein at a rank above the 50th false hit.

data did not suggest that a particular signature sparsity is optimum for achieving the best alignment between signature and jack-knifed seed proteins; however, sparsities in the range 20%–25% most commonly produce the highest specificity scores. The average alignment specificity score is 0.88 for the N-Con and ResVar signatures, that is, 88% of the key residues in the signatures are aligned to their correct (and structurally equivalent) residues in the jack-knifed seed sequences. Thus, the information within a signature is sufficient to identify family members (e.g., jack-knifed sequences) and also to consistently identify structurally equivalent key residues within the sequences of the jack-knifed proteins.

Globin-like superfamily signatures

ROC50 scores for Globin-like superfamily signatures generated using the each of the four SigGen scoring schemes (Fig. 4) range from 0.64 to 0.73; i.e., the probability of a hit known to be a member of the globin or phycocyanin family and found before the 50th false hit scoring higher than a hit known not to be related ranges between 0.64 and 0.73. This shows that information from two SCOP families can be combined into a single superfamily signature. However, signatures for the globin family achieved ROC50 scores in excess of 0.9. Inspection of the TRUE hits identified by the superfamily signatures (data not shown) showed that they were indeed detecting members from both families. Therefore the inclusion of the phycocyanin domain has two effects on the signature:

1. It has enabled the detection of phycocyanin family members. This confirms the findings of the jack-knifing experiments; that a single domain can impact significantly on signature performance.
2. The lower ROC50 scores are due to the increased variability added to the signature by the phycocyanin domain, which resulted in the detection of more false hits at ranks higher than with the globin family signature.

Nonetheless, members from both families are detected with significant ROC scores, whereas randomly generated signatures all achieve ROC50 scores of 0.0, i.e., they failed to detect any true hits from either the globin or phycocyanin families. In contrast, the globin family random signatures achieved ROC50 scores of ~0.5. We conclude that signatures derived from the Globin-like superfamily do indeed capture key features of the superfamily that have significant discriminating power. However, the performance of the signature is even more sensitive to the selection of key residues than was observed for the signatures for the single families.

Discussion

Diagnostic SCOP family signatures

We have shown that sparse signatures generated automatically for SCOP families are indeed diagnostic for their par-

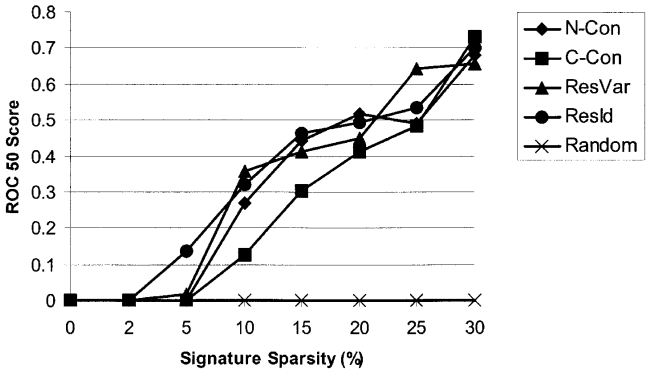

Figure 4. The graph shows the ROC50 scores plotted against signature sparsity for the N-Con, ResVar, ResId, C-Con, and random Globin-like superfamily signatures. The difference in performance between random and nonrandomly generated signatures is again obvious.

ticular family. Signatures from a range of families consistently produce ROC50 scores in excess of 0.8. In addition, signatures consistently detect more than 90% of the true known relatives identified by PSI-BLAST. This is significant considering that PSI-BLAST uses a position-specific score matrix (PSSM), incorporating all the residue information from the hits identified from a standard BLAST search, then uses the PSSM to iteratively search the database and rebuild the PSSM incorporating any new hits. Given that signatures are not iteratively refined and use the information from just a few structures to identify a subset (typically 15%–30%) of the total residues, which identify, in some cases, many hundreds of family members, the method shows considerable potential.

Signatures at the SCOP-family level were able to detect family members with less than 40% average pairwise sequence similarity to the seed proteins; this equates to an average sequence identity of 20%–30%. Therefore, signatures are able to probe relationships within the so-called "twilight zone" of sequence similarity. The Globin-like superfamily signature also highlighted the ability of a single signature to include information from two families and still maintain discriminatory power with respect to both families, while outperforming random signatures. The use of structural information from two different families introduces added flexibility into the signature which is needed to identify members of both families. However, this flexibility was the cause of reduced ROC50 scores compared to single family signatures. No individual SigGen scoring scheme consistently performed best, but random signatures consistently performed poorly. We conclude that discriminating signatures require key residues from structurally equivalent positions, and that these key residues can be automatically identified.

Potential applications of the SIGNATURE approach

Our approach has potential use in three areas of protein sequence/structure analysis. Firstly, it has been shown that relationships in the 40% sequence similarity range are identifiable. Therefore, the approach has potential as an annotation tool. To this end we have evaluated the usefulness of a database of signatures in providing annotation to protein sequences; initial results showed considerable potential and this will be described in a future work. Second is the observation that the signature-sequence alignments can identify true structural equivalences. In homology modeling, a novel sequence is aligned to sequences of known structure, and the structural features are transferred to the novel sequence. Once suitable targets are identified, the next crucial step is obtaining a high-quality alignment. Both steps can be difficult, especially for divergent proteins, yet the quality of the final model depends on them. Signatures might offer advantages over all-residue alignments because they would emphasize the biologically significant positions in the selection of the target and generation of the alignment. Finally, the residues identified provide targets for experimental investigations into the sequence–structure and structure–function relationships, and the possible role of the key residues in protein folding.

SCOP superfamily signatures

We are interested to know the generality of whether a single sparse signature can capture the features of an entire superfamily. More than 10 different superfamilies were investigated, but in none of the examples could a suitable seed alignment be generated automatically. There were too few structurally equivalent positions, and inspection of the superimpositions using molecular graphics suggested that some equivalences were incorrect. The ability to succeed only with the Globin-like superfamily reflects the broader and extremely difficult problem of making confident automated assignments of structurally equivalent residues for groups of structurally divergent proteins. Alignment methods are progressing, however, and we are exploring different alignment strategies to investigate superfamily signatures further. The use of more divergent domains in the seed alignments might also help reveal an optimum SigGen scoring scheme, because there would be less sequence similarity to "drown out" the genuinely key residues.

Development of the SIGNATURE Approach

The SIGNATURE approach is generic in the sense that signatures can in principle be generated for any family of proteins for which two or more known structures are available. We used SCOP families because they capture a high resolution of structural similarity and suggest that for many applications, a superfamily might best be represented by a collection of signatures representing the individual subgroups. For example, an unannotated sequence could be screened against a library of signatures and assigned to a broader superfamily by reference to the scores for the matches of the individual signatures comprising the superfamily. The full potential of the approach has yet to be realized, and there are several possible routes to improving the method, including the use of secondary structure information and iterative signature refinement, which will be addressed in future work.

Materials and methods

Preparation of seed alignments

Figure 5 summarizes the steps involved in generating the seed alignments. The structural alignment program STAMP was used to generate two seed alignments for each test family, one each from

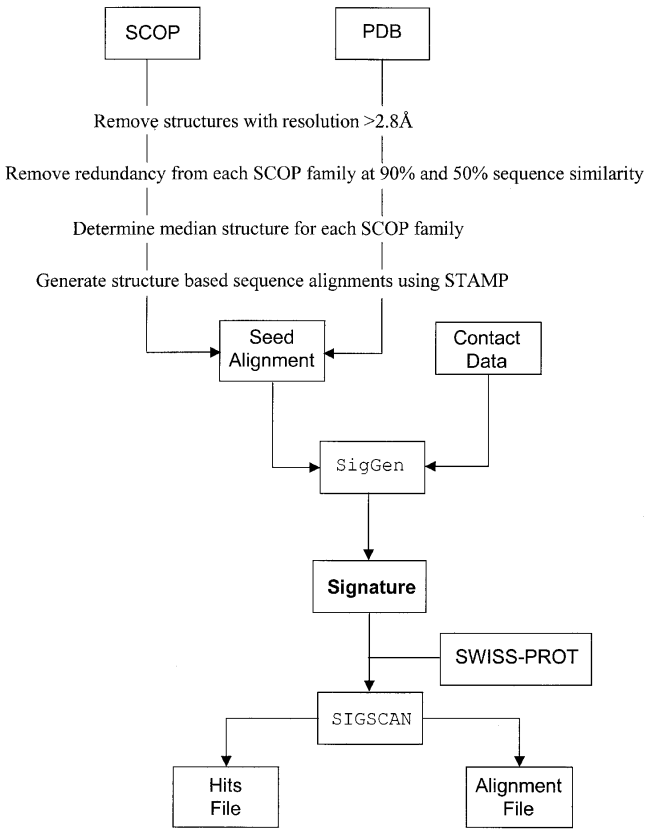

Figure 5. Summary of the approach for generating seed alignments, signature generation, and scanning. Signatures are produced from a structure-based sequence alignment (seed alignment) generated using STAMP. SigGen applies sequence-based and structure-based scoring schemes to the alignment to generate the signature; the latter requires a file of contact data describing the residue-residue contacts in the protein structure. A signature is scanned against a sequence database by using SIGSCAN and generates a hits file and alignment file for sequences returned by the search.

the DATA90 and DATA50 seed sets (see introduction). Only crystal structures of better than 2.8 Å resolution were used. STAMP uses one "scan structure" upon which all the others in the set are superimposed. The selection of the scan structure is fundamental to the quality of the resulting alignment, as shown by Gerstein and Levitt (1998), who concluded that the optimum alignment is obtained when the scan structure is the closest on average to all the other structures in the group. Accordingly we performed all possible pairwise structural alignments for a seed set and calculated the average RMSD for each structure. The scan structure has the lowest average RMSD. In the preparation of the alignment for the jack-knifed signatures (see "Validation of signatures" above), we determined a new median structure for each jack-knifed seed set.

Signature generation

SigGen selects signature positions from structurally equivalent positions in a seed alignment. Structural equivalence is indicated by a value of 1 for an alignment position in the POST_SIM output of STAMP, meaning that every pair of structures in the alignment had a P_{ij}' value for that position which is greater than a predefined threshold (6.0). The P_{ij}' value is an adaptation of the probability function of Rossmann and Argos (1976) which expresses the probability of residue structural equivalence (P_{ij}).

SigGen requires a file of residue–residue contact data (Fig. 5) derived from the three-dimensional coordinates of the seed domains. We generated each contact file using the EMBOSS application "contacts." Contact between two residues is defined as when the van der Waals surface of any atom of the first residue comes within the threshold contact distance of the van der Waals surface of any atom of the second residue. The threshold contact distance is 1 Å, and the following van der Waals radii are used: C, 1.8 Å; O, 1.4 Å; N, 1.7 Å; S, 2.0 Å; H, 1.0 Å.

SigGen scoring schemes: Residue Identity (ResId)

The ResId score for an alignment position is simply the average of all permutations of residue pair substitution scores for that position. Scores are taken from a residue substitution matrix, e.g., BLOSUM62 (Henikoff and Henikoff 1992). A ResId score is calculated for every structurally equivalent position in the seed alignment, and the average substitution scores are normalized using standard Min-Max normalization methods.

SigGen scoring schemes: Residue Variability (ResVar)

This scoring scheme implements the residue variability function of Mirny and Shakhnovich (2001):

$$s(1) = -\sum_{i=1}^{6} p_i(1) \log p_i(1) \qquad (1)$$

Where s(l) is the residue variability at position l, and $p_i(l)$ is the frequency of residues from class i at position l. Six classes of residue are defined which reflect their physicochemical properties and natural pattern of substitution:

- Aliphatic A V L I M C
- Aromatic F W Y
- Polar S T N Q
- Basic K R H
- Acidic D E
- Special G P

The special class represents the special conformational properties of glycine and proline. Because of this classification, mutations within a class are ignored—for example, L→V—whereas mutations that change the residue class are taken into account. The ResVar score for an alignment position is s(l) in equation 1. This is calculated for each structurally equivalent position and normalized as before.

SigGen scoring Schemes: Number of Contacts (N-Con)

The N-Con score is based purely on structural information: The identity and property of the residues is not considered. The N-Con score [N(l) in equation 2, below] of an alignment position reflects the number of residue–residue contacts (see above) it forms.

$$N(1) = \frac{\sum_{i=1}^{n} NC_i(1)}{n} \qquad (2)$$

Where N(l) is the average number of residue–residue contacts that the residues at position l of the seed alignment make within their respective structures. NC_i is the number of contacts that the residue (from sequence i) is involved in, and n is the number of proteins in the seed alignment. The N-Con score is calculated and normalized as before.

SigGen scoring schemes: Conservation of Contacts (C-Con)

This scoring scheme considers the number of contacts each residue at an alignment position makes, and also which residues are contacted and their position in the alignment. The score represents how conserved the contacts are. Each residue in an alignment position has associated with it a list of positions with which it makes contact. For example, if all the residues at position 25 of the seed alignment made contact with the residues at position 79 of the alignment, a conserved contact would be defined and a maximum score allocated to the residues at position 25. We repeated this procedure for all the contacts made by the residues at position 25, and an average conservation of contact score was calculated. This procedure was then repeated for every aligned position and the values normalized as before.

Key residue selection for final signature

Consider the situation where a single scoring scheme has been applied to a seed alignment containing six domains with an average sequence length of 100 residues. Each alignment position has a normalized score: 1.0 indicates the position achieved a maximum score, whereas 0.0 indicates a minimal score. The normalized scores for all structurally equivalent positions are sorted into a list in descending order. If for example a 15% sparsity signature is generated, this requires a total of 15 key residue positions to be identified. The top 15 highest normalization scores from the list are then selected. SigGen identifies the corresponding alignment position in the seed alignment and extracts the residue and gap data that will make up each key residue position of the signature. As there are six domains in the alignment in this example, each key residue position will contain a maximum of six different residue identities and six different gap values. In the case where more than one scoring scheme is to be applied in generating the signature, the normalized values from each scheme are pooled for each alignment position, and the key residue selection process then proceeds as described above.

The scoring schemes and the residue selection process described above are encapsulated in the SigGen application. SigGen has the same style command-line interface and is used in the same way as all other EMBOSS applications (see http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html for software documentation).

Signature searches

The method used for searching a sequence database with a signature is encapsulated in the EMBOSS application SIGSCAN, which implements the algorithm described in our earlier paper (Daniel et al. 1999). Previous testing of SIGSCAN (data not shown) identified several reliable parameters (Table 5) that were used for all searches. The output from SIGSCAN is a "hits file" and an "alignment file." The hits file is a list of top-scoring hits rank-ordered on the basis of score (highest-scoring hit first). The following data are given for each hit: the SWISSPROT identifier, score, rank position, start and end points of the region in the sequence to which the signature was aligned, and the classification of the hit. Classification is one of "TRUE," "FALSE," "CROSS," or "UNKNOWN" and is assigned by reference to a "gold standard" of known family members (described below). The alignment file contains the signature-sequence alignments. These show which residues in the sequences were assigned as being equivalent to the key residues contained in the signature.

Table 5. *SIGSCAN parameters used for signature searches*

	Window size	Gap extension penalty	Gap opening penalty	Hit overlap threshold	Substitution matrix used
Value	10	0.5	10	30	BLOSUM62

Reliable parameters identified from earlier testing of SIGSCAN were used for all the SIGSCAN searches. The parameters are explained in detail at http://www.hgmp.mrc.ac.uk/EMBOSS.

ROC analysis

The gold standard is a file of sequences from SWISSPROT that are uniquely related to a single SCOP family, plus sequences of ambiguous family assignment which are assigned to a SCOP superfamily or fold instead. This "validation file" was generated in a two-step process by using the SEQSEARCH and SEQSORT applications in EMBOSS, which were developed for this purpose and will be described in a publication in preparation. In brief:

1. PSIBLAST was used to search SWISSPROT with the DATA90 seed alignment for each SCOP family. The following PSIBLAST parameters were used: E-value = 0.0001, iterations = 20. An application wrapper (SEQSEARCH) to PSIBLAST was developed for this purpose.
2. The results of these searches were processed to identify both unique and overlapping hits and collated into the validation file by using SEQSORT. This file allows a classification of TRUE, CROSS, FALSE, or UNKNOWN to be assigned to each hit retrieved by a signature as follows:

- A TRUE hit is one from the same family as that from which the signature was generated.
- CROSS hits belong to a different family to that of the signature, but both belong to the same superfamily.
- FALSE hits belong to families of a SCOP fold different from that of the signature.
- UNKNOWN hits are any hits that cannot be assigned as TRUE, CROSS, or FALSE.

We generate a ROC value by calculating the area under a ROC curve truncated to the first 50 false hits:

$$\text{Area} = \frac{1}{nT} \times \sum_{i=1}^{n} Ti \qquad (3)$$

Where n is the number of false hits (i.e., 50), T is the total number of known family members taken from the validation file, and Ti is the number of TRUE hits detected above the ith FALSE hit. Thus to calculate ROC50, Ti is summed from i = 1 to i = 50.

Alignment specificity scores

Consider a signature containing 10 key residue positions. When that signature is scanned against a sequence from SWISSPROT, an alignment is produced in which each of the 10 key residue positions is aligned to a single residue in the SWISSPROT sequence. An alignment specificity score of 1.0 means that each of the 10 residues identified in the sequence are structurally equivalent to the position in the seed set alignment from which the matching key residue position was derived. An alignment specificity score of 0.40 means that four out of the 10 residues in the sequence are structurally equivalent to the key residue positions in the original seed set alignment. The alignment specificity for each alignment of a jack-knifed signature to the sequence that was jack-knifed out was calculated as follows:

$$\text{Alignment specificity} = \frac{\text{No. of correctly aligned residues in jack-knifed sequence}}{\text{Total no. of key residue positions in the signature}} \quad (4)$$

A correctly aligned residue (X) is one that is structurally equivalent to the key residue position (Y) in question; if the jack-knifed domain was put back into the seed alignment and the signature regenerated, then residue X would form part of key residue position Y.

Acknowledgments

M.J.B. was funded by a BBSRC CASE Award studentship sponsored by Novo Nordisk, Denmark. J.C.I and R.R. were funded by the MRC. We thank Dr. Robert Bywater (formerly of Novo Nordisk, Denmark) for many hours of useful discussions and help with the manuscript.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W., and Willett, P. 1994. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **243:** 327–344.

Blake, J.D. and Cohen, F.E. 2001. Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* **307:** 721–735.

Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., and Sauer, R.T. 1990. Deciphering the message in protein sequences: Tolerance to amino acid substitutions. *Science* **247:** 1306–1310.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Daniel, S.C., Parish, J.H., Ison, J.C., Blades, M.J., and Findlay, J.B. 1999. Alignment of a sparse protein signature with protein sequences: Application to fold prediction for three small globulins. *FEBS Lett.* **459:** 349–352.

Dosztanyi, Z., Fiser, A., and Simon, I. 1997. Stabilization centers in proteins: Identification, characterization and predictions. *J. Mol. Biol.* **272:** 597–612.

Friedberg, I. and Margalit, H. 2002. Persistently conserved positions in structurally similar, sequence dissimilar proteins: Roles in preserving protein fold and function. *Protein Sci.* **11:** 350–360.

Gerstein, M. and Levitt, M. 1998. Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins. *Protein Sci.* **7:** 445–456.

Gribskov, M. and Robinson, N.L. 1996. Use of receiver operating characteristics (ROC) analysis to evaluate sequence matching. *Comp. Chem.* **20:** 25–33.

Hargbo, J. and Elofsson, A. 1999. Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins* **36:** 68–76.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Ison, J.C., Blades, M.J., Bleasby, A.J., Daniel, S.C., Parish, J.H., and Findlay, J.B. 2000. Key residues approach to the definition of protein families and analysis of sparse family signatures. *Proteins* **40:** 330–341.

Jennings, A.J., Edge, C.M., and Sternberg, M.J. 2001. An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Eng.* **14:** 227–231.

Kannan, N. and Vishveshwara, S. 1999. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.* **292:** 441–464.

Kannan, N., Selvaraj, S., Gromiha, M.M., and Vishveshwara, S. 2001. Clusters in α/β barrel proteins: Implications for protein structure, function, and folding: A graph theoretical approach. *Proteins* **43:** 103–112.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14:** 846–856.

Kleywegt, G.J. 1999. Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **285:** 1887–1897.

Li, W.W., Reddy, B.V., Tate, J.G., Shindyalov, I.N., and Bourne, P.E. 2002. CKAAPs DB: A conserved key amino acid positions database. *Nucleic Acids Res.* **30:** 409–411.

Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. 2000. SCOP: A structural classification of proteins database. *Nucleic Acids Res.* **28:** 257–259.

Markiewicz, P., Kleina, L.G., Cruz, C., Ehret, S., and Miller, J.H. 1994. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J. Mol. Biol.* **240:** 421–433.

Milla, M.E., Brown, B.M., and Sauer, R.T. 1994. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nat. Struct. Biol.* **1:** 518–523.

Mirny, L. and Shakhnovich, E. 2001. Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308:** 123–129.

Mirny, L.A., Abkevich, V.I., and Shakhnovich, E.I. 1998. How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci.* **95:** 4976–4981.

Orengo, C.A. 1999. CORA—Topological fingerprints for protein structural families. *Protein Sci.* **8:** 699–715.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284:** 1201–1210.

Reddy, B.V., Li, W.W., Shindyalov, I.N., and Bourne, P.E. 2001. Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. *Proteins* **42:** 148–163.

Rossmann, M.G. and Argos, P. 1976. Exploring structural homology of proteins. *J. Mol. Biol.* **105:** 75–95.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng.* **12:** 85–94.

Russell, R.B. and Barton, G.J. 1992. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins* **14:** 309–323.

Shakhnovich, E., Abkevich, V., and Ptitsyn, O. 1996. Conserved residues and the mechanism of protein folding. *Nature* **379:** 96–98.

Spang, R. and Vingron, M. 2001. Limits of homology detection by pairwise sequence comparison. *Bioinformatics* **17:** 338–342.

Suckow, J., Markiewicz, P., Kleina, L.G., Miller, J., Kisters-Woike, B., and Muller-Hill, B. 1996. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J. Mol. Biol.* **261:** 509–523.